Publishing and Using Plant Names as an Ontology Service

Jouni Tuominen, Nina Laurenne, and Eero Hyvönen

Semantic Computing Research Group (SeCo)
Aalto University School of Science, Dept. of Media Technology, and
University of Helsinki, Dept. of Computer Science
http://www.seco.tkk.fi, firstname.lastname@aalto.fi

Abstract. Animals and plants are referred to using scientific or common names depending on the expertise of an audience or a source of data. The names change in time and therefore their usage as identifiers as such is problematic. We present a solution for managing and using plant names as an ontology. The ontology is based on the TaxMeOn meta-ontology for biological names. In order to refer to organisms unambiguously and publish information as Linked Data on the web, the names are given URIs. The ontology is developed collaboratively and it supports the approval process and temporal tracking of the common names. We introduce an ontology service of plant names for end-users and provide user interfaces and APIs for integrating the ontology into applications.

1 Introduction

The scientific names of plants and animals have a major role when indexing, querying, and integrating information about species. Biologists use scientific names although the vast majority of people use the common name equivalents. Contrary to common belief, neither the scientific nor common names identify organisms unambiguously as one name may point to multiple species and one species may have multiple names.

New research results change the name combination of the scientific names because taxa are constantly split and lumped. For example, if a species is changed into another genus, the name combination changes accordingly. Approximately 25,000 new species descriptions are published in thousands of journals annually [6] which makes it hard for researchers to keep up-to-date the biodiversity of the nature. Not all organisms need a common name but still there is huge work to be done in developing the vernacular nomenclature and in terms of established names, the dialect expressions remarkably expand the spectrum of the biological names.

The international commissions of the nomenclatures (IBC, ICZN) specify the rules how the scientific names should be used in various taxonomic treatments. The nomenclatures of plants and animals are independent of each other and the rules are applied only to the scientific names. The common names are not regulated but they also change in time because there is often a need to update the common names at intervals. The changing nature of the names poses challenges for their management [5, 10, 13].

The diversity of the names causes problems when combining data from heterogeneous sources, e.g., observational records, literature and museum collections [11, 9]. The data cannot be easily integrated if a taxon is referred to using multiple names and vice versa the existence of homonyms (the same name refers to multiple taxa) causes errors when merging the data.

Comprehensive reference lists and catalogues of the names have been proposed as a solution to facilitate the access to the names [1, 10]. However, this is not enough because the biological names ought to be machine-processable in order to refer to them unambiguously and semantically enrich the biological contents. Ontologies remarkably increase the re-use and utilization of the available resources which minimizes the amount of manual work when harmonizing data.

We present an ontology model for managing the common names of organisms and linking them to the scientific names. The model supports temporal tracking of name changes and an approval process of the common names. The model is used for maintaining and publishing plant names in Finnish as an ontology. The ontology is published as Linked Open Data [3] and can be used as an ontology service.

2 Ontology Model

TaxMeOn[1] [14] is an RDF-based meta-ontology for modeling and managing biological names and classifications. TaxMeOn introduces classes and properties for expressing biological names as ontologies. The model consists of three parts according to the level of taxonomic details, which are common names, species checklists, and detailed taxonomic information respectively. In this paper, the focus is on the common names although many of the classes and properties are common to all three parts. The simplified structure of the model is presented in Fig. 1, where the core classes are *Scientific name*, *Common name* and their statuses. The status of the *Scientific name* indicates if a name is an accepted or a synonymized one, etc. The synonyms are linked to an accepted name. The hierarchical structure is constructed setting relations between the *Scientific names.*

The *Common names* (in one or more languages) that refer to the same taxon are connected through a *Scientific name.* The model also allows mapping the scientific names to each other based on the underlying taxonomic concepts (congruence, overlap, part-of, general association). A taxonomic rank expresses the hierarchical level in a classification (e.g., a species, a genus) and it is specified for every scientific name. The taxonomic ranks are presented as a separate vocabulary which contains 61 ranks, of which 60 are obtained from TDWG Taxon Rank LSID Ontology[2]. In order to avoid the complex details of the botanical and

[1] `http://schema.onki.fi/taxmeon/`

[2] `http://rs.tdwg.org/ontology/voc/TaxonRank`

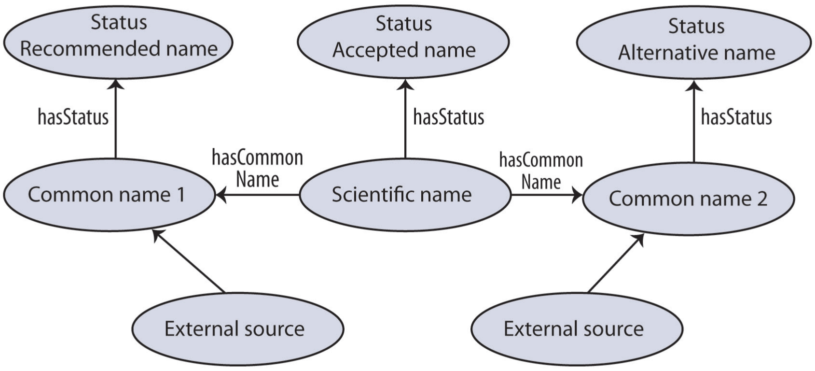

Fig. 1. The ontology model of the common names of organisms. The ellipses represent classes and the arrows depict relations between the classes.

horticultural nomenclatures, the species level and the taxonomic levels below it are treated as one unit.

The approval process of the common names is the following: first, a new name is proposed; then the name becomes accepted; and finally, the name may become an alternative, if there is a new accepted name for the same plant. The model allows the maintainers to propose a new name which then can be commented by the other maintainers until the name finally gets accepted, rejected or synonymized. The temporal management of the names is based on time stamps which are given to the statuses of the names in the approval process. If a name is given a new status, the old status is not removed from the system. This makes it possible to track the chain of changes of the names and to see the period of time period when a particular name was accepted.

3 Managing Plant Names as an Ontology

We applied the TaxMeOn ontology model to a database of the Finnish names of plants maintained by the Finnish Biology Society Vanamo[3]. The original database contained nearly 26,000 plant names in Finnish in a single classification. The taxa were divided into three taxonomic levels (a species, a genus and a family) but it is possible to specify more taxonomic levels in the current ontology.

The database of the plant names was converted into RDF format based on the TaxMeOn ontology model. The ontology is managed in the metadata editor SAHA[4] [7] by the Vanamo association. Currently, the ontology contains 21,797 species, and the number of updates exceeds one thousand names yearly. The

[3] http://www.vanamo.fi

[4] http://www.seco.tkk.fi/services/saha/

utilization of the ontology facilitates the management of the names because the approval process is integrated into the ontology.

The association has an active role in developing new Finnish names for plants and the public availability of the ontology releases voluntary based work for more relevant activities than responding to various queries by journalists, translators etc. The development of the new names is based on the needs, therefore the coverage of the taxa is not systematically or geographically restricted into any particular plant group or a region.

The browser-based SAHA editor allows collaborative editing of the ontology, providing the simultaneous access of multiple users and a chat functionality. The TaxMeOn model has been extended to support the management of the ontology in SAHA, by adding a property indicating the current status of the processing of a proposed common name. If a new name is suggested for a species, a maintainer can add it into the ontology and mark it as “in process”. The proposed but not yet processed names can be found easily at later stages of the process.

4 Using Plant Names as an Ontology Service

The ontology is published as Linked Open Data in the Finnish Ontology Library Service ONKI[5] [15], as part of the Finnish semantic web infrastructure project FinnONTO[6] [4]. The ONKI service provides user interfaces and APIs for accessing and using the plant names in applications. For example, end-users can browse and search the ontology to find a common name for a taxon that they know only by the scientific name. The ONKI selector widget can be integrated into legacy CMS systems to provide an autocomplete and URI fetching features to support the annotation of plant related information.

One of the advantages of the ontology service is that the end-users can now access the ontology themselves. Users are directed to the ONKI service via search engines, and they have adopted the service by extending Wikipedia articles about plant species with links to Finnish plant names in ONKI. End-users actively send feedback, comments and corrections to the maintainers, which help them to improve the quality of the content.

The ontology is also accessible as a SPARQL endpoint. An example query below shows how the accepted Finnish common names of species (and taxa below it) that belong to a genus “*Quercus*” (oak) can be retrieved:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX taxmeon: <http://www.yso.fi/onto/taxmeon/>
PREFIX taxonomic-ranks: <http://www.yso.fi/onto/taxonomic-ranks/>

SELECT ?vernacularName WHERE {
  ?species taxmeon:isPartOfHigherTaxon ?genus .
  ?genus rdf:type taxonomic-ranks:Genus .
  ?genus rdfs:label "Quercus"^^xsd:string .
```

[5] http://onki.fi/en/browser/overview/kassu

[6] http://www.seco.tkk.fi/projects/finnonto/

```
  ?species taxmeon:hasVernacularName ?vernacularNameRes .
  ?vernacularNameRes taxmeon:hasVernacularNameStatus ?status .
  ?status rdf:type taxmeon:AcceptedVernacularName .
  ?vernacularNameRes rdfs:label ?vernacularName .
  FILTER langMatches(lang(?vernacularName), "fi")
}
```

The result of the query is a list of the Finnish names of oak species, such as the sessile oak and white oak. The query demonstrates the use of the ontology for cross-language query expansion.

Currently, the plant name ontology is used by several cultural museums and libraries for annotating their collections. The ontology is also applied as a use case in the EU FP project ENVIROFI[7] which focuses on the environmental usage area of the Future Internet. The ontology is used in the project as a conceptual hub for referring to the plants in the observational data on biodiversity. The ontology has been extended with the English and German names of plants used in the project pilots (these names are not available in the ONKI ontology service).

5 Discussion

5.1 Related Work

The importance of persistent identifiers for organism names and solutions for managing them on the semantic web have been discussed by several workers. Page [8] presented how taxon names are modeled as semantic metadata in RDF form. Taxon names are identified with using Life Science Identifiers (LSID) and the names are connected using taxonomic relations. Taxon names that are obtained from various data sources and which refer to the same taxon are mapped using the *owl:sameAs* relation. Schulz et al. [12] presented the first ontology model of biological taxa and its application to physical individuals. The model is based on a single unchangeable classification. Franz and Thau [2] evaluated the limitations of applying ontologies to the scientific names and concluded that ontologies should focus either on a nomenclatural point of view or on strategies for aligning multiple taxonomies.

The Darwin Core (DwC)[8] is a metadata schema developed for taxon occurrence data by the TDWG (Biodiversity Information Standards). The goal of DwC is to standardize the form of how biological information is presented. However, it lacks the semantic aspect and when it comes to the names, the scope of DwC is quite general.

Taxonconcept.org[9] provides Linked Open Data identifiers for species concepts and links data from different sources. All the names of species are expressed using literals. Also, the machine-processability is weakened by the usage of literal values for expressing the hierarchies. The data contains scientific and common names, and taxonomic statuses.

[7] http://www.envirofi.eu

[8] http://www.tdwg.org/standards/450/

[9] http://www.taxonconcept.org

Many existing databases aim to be comprehensive online catalogues that aggregate individual species checklists, such as the Catalogue of Life (CoL)[10] and The International Plant Names Index (IPNI)[11]. The IPNI database contains only scientific names, but the Catalogue of Life also includes their taxonomic statuses and common names. They both provide the names in a machine-processable form, as RDF conforming to the TDWG Taxonomic Concept Transfer Schema (TCS)[12] using LSIDs as identifiers of the names [5]. In the Catalogue of Life the requirement to use a separate LSID resolver for fetching metadata about an LSID prevents the Linked Data compatibility of the dataset. The IPNI database provides an LSID proxy that allows Linked Data compatibility. In the IPNI database, the hierarchy is not expressed explicitly in the RDF (e.g., the genus of a species is shown only in the binomial name literal).

There are several other plant name databases available on the web, e.g., the Royal Horticultural Society Horticultural Database[13], The Plant List[14] and the Euro+Med PlantBase[15]. Most available resources contain the scientific names, but in few, the common names are included. Common to these systems is that they are intended for human usage, and they are not available in a machine-processable form with unique name identifiers.

5.2 Contributions and Future Work

Most of the related work concentrate on the scientific names, but our focus is on the common names. The common names expand the cross-domain use of the ontology because they are in wider spectrum of use than the scientific ones. The ontology is available in machine-processable RDF format, with explicit semantics, e.g., the hierarchical relations are set between the plant URIs, and the statuses of names are supported. The TaxMeOn model provides a solution for managing the approval process of common names, supporting the temporal tracking of the name changes via statuses and their time stamps. The model connects together different names of a taxon facilitating data integration and information retrieval in cases where data is combined from heterogeneous sources.

We have also demonstrated the complete workflow from a collaborative development of an ontology to publishing it as Linked Open Data and as an ontology service which makes it accessible to the general public. The plant name ontology helps harmonizing the terminology which in turn enhances communication between various users. Application developers can utilize the ontology by using the plant name URIs for unambiguous referencing to plants species.

Currently, hybrid taxa are modeled in the ontology in the same way as the ordinary species. An idea for the future development is to extend the model to

[10] http://www.catalogueoflife.org

[11] http://www.ipni.org

[12] http://www.tdwg.org/standards/117/

[13] http://apps.rhs.org.uk/horticulturaldatabase

[14] http://www.theplantlist.org

[15] http://www.emplantbase.org

support the representation of hybrid names at a deeper level. Another area for development is to link the scientific names of plants to their author URIs in DBpedia, connecting the ontology to the Linked Data Cloud (LOD).

Ontologies are a bridge between experts and ordinary people in communication and popularizing science. Additionally, the Linked Data approach provides a way how to easily extend an ontology with additional information which in turn increases the information value of contents.

Acknowledgments This work is part of the National Semantic Web Ontology project in Finland FinnONTO (2003-2012), funded mainly by the National Technology and Innovation Agency (Tekes) and a consortium of 38 public organizations and companies, and the EU FP project The Environmental Observation Web and its Service Applications within the Future Internet (ENVIROFI). We thank Leo Junikka and Arto Kurtto for their collaboration.

References

1. Dengler, J., Berendsohn, W.G., Bergmeier, E., Chytrý, M., Danihelka, J., Jansen, F., Kusber, W.H., Landucci, F., Müller, A., Panfili, E., Schaminée, J.H.J., Venanzoni, R., von Raab-Straube, E.: The need for and the requirements of EuroSL, an electronic taxonomic reference list of all european plants. Biodiversity & Ecology 4, 15–24 (2012)
2. Franz, N., Thau, D.: Biological taxonomy and ontology development: scope and limitations. Biodiversity Informatics 7, 45–66 (2010)
3. Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1–136, Morgan & Claypool (2011)
4. Hyvönen, E., Viljanen, K., Tuominen, J., Seppälä, K.: Building a national semantic web ontology and ontology service infrastructure—the FinnONTO approach. In: Proceedings of the ESWC 2008, Tenerife, Spain. Springer–Verlag (2008)
5. Jones, A.C., White, R.J., Orme, E.R.: Identifying and relating biological concepts in the Catalogue of Life. Biomedical Semantics 2(7) (2011)
6. Knapp, S., Polaszek, A., Watson, M.: Spreading the word. Nature 446, 261–262 (2007)
7. Kurki, J., Hyvönen, E.: Collaborative metadata editor integrated with ontology services and faceted portals. In: Workshop on Ontology Repositories and Editors for the Semantic Web (ORES 2010), the Extended Semantic Web Conference ESWC 2010, Heraklion, Greece. CEUR Workshop Proceedings, `http://ceur-ws.org` (2010)
8. Page, R.: Taxonomic names, metadata, and the semantic web. Biodiversity Informatics 3, 1–15 (2006)
9. Page, R.D.M.: Biodiversity informatics: the challenge of linking data and the role of shared identifiers. Briefings in Bioinformatics 9(5), 345–354 (2008)
10. Patterson, D.J., Cooper, J., Kirk, P.M., Pyle, R.L., Remsen, D.P.: Names are key to the big new biology. Trends in Ecology & Evolution 25(12), 686–691 (2010)
11. Sarkar, I.N.: Biodiversity informatics: organizing and linking information across the spectrum of life. Briefings in Bioinformatics 8(5), 347–357 (2007)

12. Schulz, S., Stenzhorn, H., Boeker, M.: The ontology of biological taxa. Bioinformatics 24(13), 313–321 (2008)
13. Segers, H., de Smet, W.H., Fischer, C., Fontaneto, D., Michaloudi, E., Wallace, R.L., Jersabek, C.D.: Towards a list of available names in zoology, partim phylum rotifera. Zootaxa 3179, 61–68 (2012)
14. Tuominen, J., Laurenne, N., Hyvönen, E.: Biological names and taxonomies on the semantic web – managing the change in scientific conception. In: Proceedings of the ESWC 2011, Heraklion, Greece. pp. 255–269. Springer–Verlag (2011)
15. Viljanen, K., Tuominen, J., Hyvönen, E.: Ontology libraries for production use: The Finnish ontology library service ONKI. In: Proceedings of the ESWC 2009, Heraklion, Greece. pp. 781–795. Springer–Verlag (2009)